Massively scalable Sinkhorn distances via the Nyström method

Sebastiano Scardera

Presentation of the problem

Let $\mathbf{p}, \mathbf{q} \in \Delta_n$, the Sinkhorn distance with parameter $\eta > 0$ is defined as

$$W_\eta(\mathbf{p}, \mathbf{q}) := \min_{P \in \mathcal{M}(\mathbf{p},\mathbf{q})} \sum_{i,j} P_{ij} \left\|x_i - x_j\right\|_2^2 - \eta^{-1} H(P) = \min_{P \in \mathcal{M}(\mathbf{p},\mathbf{q})} V_C(P),$$

where $H(P) := \sum_{i,j} P_{ij} \log \frac{1}{P_{ij}}$. In the course, we saw that

$$W_\eta(\mathbf{p}, \mathbf{q}) = \min_{P \in \mathcal{M}(\mathbf{p},\mathbf{q})} \eta^{-1} \mathrm{KL}\left(P \| \mathrm{K} \odot (\mathbf{p} \otimes \mathbf{q})\right), \quad \mathrm{K}_{ij} := e^{-\eta \left\|x_i - x_j\right\|^2}.$$

Objective: Approximate $W_\eta(\mathbf{p}, \mathbf{q})$ in an efficient way.

Proposed method

Idea: Run SINKHORN on low-rank approximation kernel

NYS-SINK: Nyström method + Sinkhorn algorithm

- ADAPTIVENYSTRÖM: searches for a good rank-lowest possible Nyström approximation.
- SINKHORN: is the classical version computed on the approximated kernel.
- ROUND: starting from Sinkhorn's result, it returns a feasible solution, not too far from the original matrix.

Preliminaries

Definition 1 (Effective dimension)

Let $\lambda_j(K)$ the jth largest eignevalue of K. The *effective dimension* of K at level $\tau > 0$

$$d_{\text{eff}}(\tau) := \sum_{j=1}^{n} \frac{\lambda_j(K)}{\lambda_j(K) + \tau n}.$$

Definition 2 (Approximation rank)

Given $X = \{x_1, ..., x_n\} \subseteq \mathbb{R}^d$ with $\|x_i\| \leq R$ for all $i \in [n]$, $\eta > 0$, and $\epsilon' \in (0, 1)$, the *approximation rank* is

$$r^*(X, \eta, \epsilon') := d_{\text{eff}}\left(\frac{\epsilon'}{2n} e^{-4\eta R^2}\right),$$

where $d_{\text{eff}}(\cdot)$ is the effective rank of the kernel matrix $K := e^{-\eta C}$.

Main result

Theorem 1

Let $\epsilon, \delta \in (0,1)$. NYS-SINK runs in $\tilde{O}\left(nr\left(r + \frac{\eta R^4}{\epsilon}\right)\right)$ time, uses O(n(r+ d)) space, and returns a feasible matrix $\hat{P} \in \mathcal{M}(\mathbf{p}, \mathbf{q})$ in factored form and $r \in \mathbb{N}$, where

$$\left| V_C(\hat{P}) - W_\eta(\mathbf{p}, \mathbf{q}) \right| \leq \epsilon \quad \text{and} \quad \mathrm{KL}\left(\hat{P} \| P^\eta\right) \leq \eta\epsilon$$

and, with probability 1 - δ,

$$r \leq c \cdot r^* \left(X, \eta, \epsilon'\right) \log \frac{n}{\delta},$$

for a universal constant c and where $\epsilon' = \Omega\left(\epsilon R^{-2}\right)$.

Adaptability

Corollary 3

If X lies on a suitable k-dimensional manifold, then with high probability

$$d_{eff}(\tau) \leq \left(c_1 \log \frac{1}{\tau}\right)^{5k/2} + c_2 \quad \textit{and so} \quad r^*(X, \eta, \epsilon') \leq c_{\Omega,\eta} \left(\log \frac{n}{\epsilon'}\right)^{5k/2}.$$

As a consequence NYS-SINK *requires* $\tilde{O}\left(n \cdot \frac{c_{\Omega,\eta}}{\epsilon} \left(\log \frac{n}{\epsilon}\right)^{5k}\right)$ *time.*

Key steps

Lemma 1 (Nyström approximation of Gaussian kernel)

Let $(\tilde{K}, r)$ output of $\textsc{AdaptiveNyström}(X, \eta, \tau)$. Then:

$$\left\|K - \tilde{K}\right\|_\infty \leq \tau \quad \text{and} \quad \mathbb{P}\left(r \leq c \cdot d_{\text{eff}}\left(\frac{\tau}{n}\right) \log\left(\frac{n}{\delta}\right)\right) \geq 1 - \delta.$$

Theorem 2 (Stability of Sinkhorn projections)

If $K = e^{-\eta C}$ and if $\tilde{K} \in \mathbb{R}_{>0}^{n \times n}$ satisfies $\left\|\log K - \log \tilde{K}\right\|_\infty \leq \epsilon'$, then

$$\left\|\tilde{P}\mathbf{1} - \mathbf{p}\right\|_1 + \left\|\tilde{P}^T\mathbf{1} - \mathbf{q}\right\|_1 \leq \epsilon' \quad \text{and} \quad \left|V_C(P^\eta) - V_C(\tilde{P})\right| \leq \frac{\epsilon}{2}$$

where $\tilde{P} := D_1 \tilde{K} D_2$ and $D1, D2$ are the outputs of $\textsc{Sinkhorn}$.

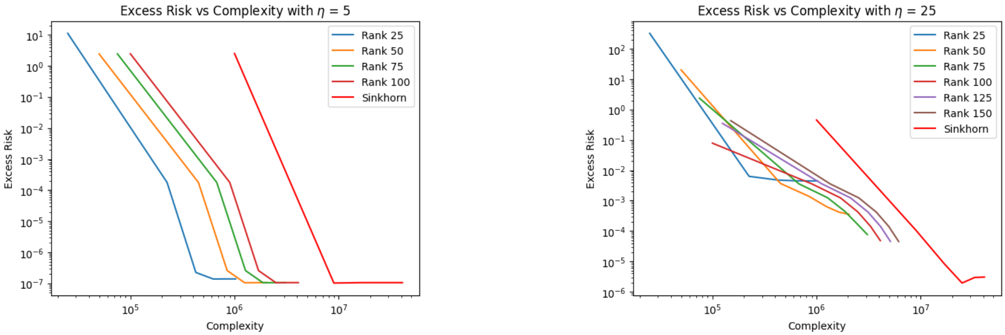
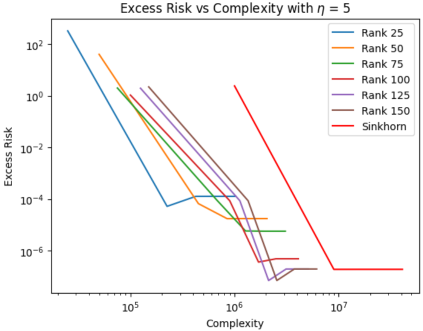


Figure: Time-accuracy tradeoff for NYS-SINK and SINKHORN in dimension 2, for a range of regularization parameters and approximation ranks r

Experimental results



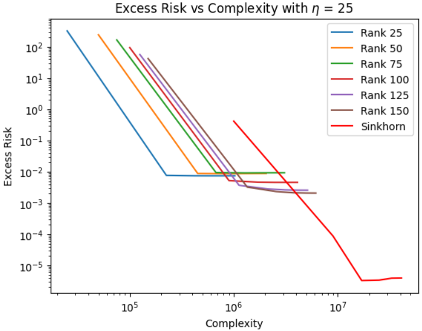


Figure: Time-accuracy tradeoff for NYS-SINK and SINKHORN in dimension 3, for a range of regularization parameters and approximation ranks r

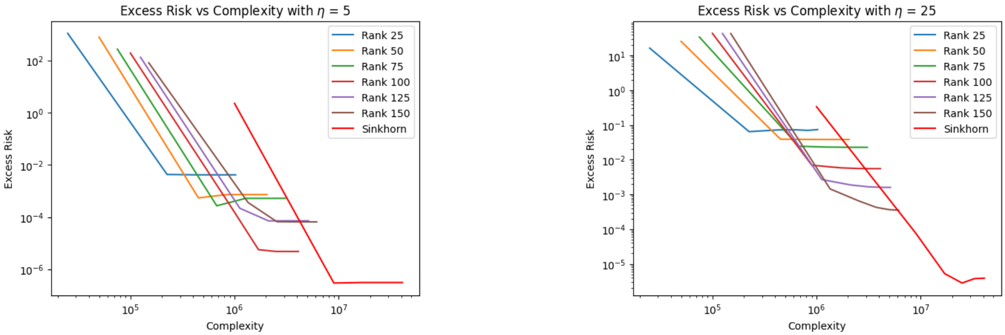


Figure: Time-accuracy tradeoff for NYS-SINK and SINKHORN in dimension 5, for a range of regularization parameters and approximation ranks r

Critics

- NYS-SINK convergence results valid only for the squared Euclidean distance (need for Gaussian kernel);
- Numerical issues related to the positive definiteness of the approximated kernel;
- No numerical test with the adaptive algorithm (only fixed rank and number of iterations).

Conclusion and future perspective

NYS-SINK:

- Fast, reliable and adaptive
- New theoretical guarantees on low-rank Sinkhorn on Gaussian kernels
- Independent interesting results on Nyström Gaussian kernel approximation and Sinkhorn projections stability

Future works:

- Approximate a broader class of kernels
- Sharper constants for the bounds
- Understanding why NYS-SINK performs well even when r is smaller than the theoretical guarantees.